Comparative Responsiveness of the PROMIS Pain Interference Short Forms with Legacy Pain Measures: Results from Three Randomized Clinical Trials

Chen X. Chen,[1] Kurt Kroenke,[2,3,4] Timothy Stump,[5] Jacob Kean,[6,7] Erin E. Krebs,[8,9] Matthew J. Bair,[2,3,4] Teresa Damush,[2,3,4,10] Patrick O. Monahan[5]

[1]Indiana University School of Nursing, Indianapolis, IN, USA;

[2]Indiana University School of Medicine, Indianapolis, IN, USA;

[3]Regenstrief Institute, Inc. Indianapolis, IN, USA;

[4]VA Health Services Research and Development Center for Health Information and Communication, Indianapolis, IN, USA;

[5]Indiana University School of Medicine, Department of Biostatistics, Indianapolis, IN, USA;

[6]University of Utah School of Medicine Department of Population Health Sciences, Salt Lake City, UT, USA;

[7]Salt Lake VA Health Care System Decision-Enhancement and Analytic Sciences Center, Salt Lake City, UT, USA;

[8]Minneapolis VA Center for Care Delivery and Outcomes Research, Minneapolis, MN, USA;

[9]University of Minnesota Medical School, Minneapolis, MN, USA;

[10]VA Health Services Research and Development, Precision Monitoring for Quality Improvement (PRIS-M QUERI Center), Indianapolis, IN, USA

Corresponding Author: Chen X. Chen, 600 Barnhill Drive E415, Indianapolis, IN 46202

Email: cxchen@iu.edu, Tel: (317) 274-7441

Conflict of Interest: None.

Disclosures: This work was supported by a National Institute of Arthritis and Musculoskeletal Disorders R01 award to Dr. Monahan (R01 AR064081) and Department of Veterans Affairs Health Services Research and Development Merit Review awards to Drs. Bair (IIR 10-128), Krebs (IIR 11-125), and Damush (VA HSRD QUERI Service Directed Project SDP- 10-379). Dr. Chen was supported by the National Institute of Nursing Research under award number 5T32 NR007066, the Indiana University–Purdue University Indianapolis Enhanced Mentoring Program with Opportunities for Ways to Excel in Research (EMPOWER) Grant, and Grants Numbers KL2TR001106 and UL1TR001108 (Shekhar, PI) funded by the National Institutes of Health, National Center for Advancing Translational Sciences Clinical and Translational Sciences Award. Dr. Kean was supported by the Department of Veterans Affairs Rehabilitation Research and Development Career Development Award (IK2RX000879). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Department of Veteran Affairs.

- The PROMIS Pain Interference short forms are responsive to change.
- Their responsiveness are comparable to legacy pain measures.
- Responsiveness may vary based on the sample and the direction of change.

ABSTRACT

The PROMIS Pain Interference (PROMIS-PI) scales are reliable and publicly accessible; however, little is known about how responsive they are to detect change in clinical trials and how their responsiveness compares to legacy measures. The study purpose was to evaluate responsiveness for the PROMIS-PI scales and to compare their responsiveness with legacy pain measures. We used data from three clinical trials totaling 759 participants. The clinical trials included patients with chronic low back pain (n= 261), chronic back or osteoarthritis pain (n

= 240), and a history of stroke (n= 258). At both baseline and follow-up, participants completed PROMIS-PI scales and legacy pain measures (Brief Pain Inventory Interference scale, Pain/Enjoyment/General Activity (PEG) scale, SF-36 Bodily Pain scale, and Roland-Morris Disability Questionnaire). We measured global ratings of pain change, both prospectively and retrospectively, as anchors to identify patients as improved, unchanged, or worsened. Responsiveness was assessed with standardized response means, statistical tests comparing change groups, and area-under-curve analysis. The PROMIS-PI scales had largely comparable responsiveness with the Brief Pain Inventory Interference scale and PEG. The four PROMIS-PI short forms had comparable responsiveness. For all pain questionnaires, responsiveness varied based on the study population and whether pain improved or worsened.

Perspective: This paper presents (1) how responsive the PROMIS Pain Interference scales were to detect change over time in the context of three clinical trials and (2) how their responsiveness compared to legacy pain measures. The findings can help researchers and clinicians choose between different patient-reported pain outcome measures.

Keywords: Pain Interference; Pain Measurement; PROMIS; Responsiveness; Patient-reported outcome measures

1. Introduction

The National Institutes of Health-funded Patient-Reported Outcomes Measurement Information System (PROMIS®) provides researchers and clinicians with outcome measures that are reliable, valid, and publicly accessible.[1; 3] In contrast to legacy measures, PROMIS® emerged from state-of-art psychometric methodologies including item response theory.[3] Item response theory supports computerized adaptive testing and development of fixed-length short forms containing the most informative items.[1]

One key PROMIS® domain is Pain Interference, because pain interference with daily activities is a recognized core outcome in pain research and clinical care.[1; 8] The PROMIS Pain Interference (PROMIS-PI) scales measure the degree to which pain interferes with physical, emotional, and social activities, and are available for adults, pediatric self-report, and parent proxy-report. The PROMIS-PI scales for adults are the focus of this study. Four fixed-length adult PROMIS-PI short forms are available: one with 4 items, two with 6 items, and one with 8 items.

Although PROMIS-PI short forms have advantages of brevity, public accessibility, and high precision, they are relatively new. Legacy measures–including the Brief Pain Inventory (BPI),[5] Pain, Enjoyment of Life, General Activity (PEG) scale,[17] SF-36 Bodily Pain subscale, and Roland-Morris Disability Questionnaire–have been more extensively evaluated in clinical populations. It is particularly warranted to evaluate PROMIS-PI's responsiveness.

Responsiveness focuses on a measure's ability to detect changes over time.[25] Researchers sometimes use "responsiveness" interchangeably with "sensitivity to change". In this paper, we use the term responsiveness to prevent potential confusion with the term "sensitivity" used in diagnostic research.[24] Evidence of responsiveness is vital for cohort studies and clinical trials of efficacy/effectiveness. Using measures with good responsiveness minimizes the risk of false negative trials and reduces sample size requirements.

The evidence is limited and mixed about PROMIS-PI responsiveness. Responsiveness of PROMIS-PI was supported in a few studies.[30,7,2] These studies revealed significant changes in PROMIS-PI scores post-treatment [2; 29; 30] or expected relationships between changes in PROMIS-PI and anchor measures.[7; 21] In other studies, however, the responsiveness of PROMIS-PI was not well supported. In one study,[12] the PROMIS-PI scale change scores did not differ between pain improved and non-improved groups. In another study,[14] the PROMIS-PI was less responsive than the legacy measures (BPI and PEG).[14]

Previous studies are limited in three ways. First, most studies [2; 7; 12; 21; 29; 30] did not compare responsiveness of the PROMIS-PI scales with legacy measures, which precluded conclusions regarding relative superiority of measures. Second, five out of six studies [2; 7; 12; 29; 30] assessed only one version of the PROMIS-PI scales, which precluded head-to-head comparisons between PROMIS-PI short forms with different lengths. Third, except for two studies,[14; 21] PROMIS-PI short forms were not evaluated in the context of clinical trials. More evidence on PROMIS-PI's responsiveness in clinical trials is required before researchers can be confident in adopting them as a primary trial outcome.[2]

Given the limitations of previous studies, our study purpose was to evaluate responsiveness for the fixed-length PROMIS-PI short forms. We evaluated the four PROMIS-PI short forms in the context of three clinical trials, and compared responsiveness of the PROMIS-PI short forms to legacy measures.

2. Methods

2.1 Design and Participants

Data from a total sample of 759 patients were analyzed in this study. Participants were recruited from 3 participating clinical trials.

Sample 1 included 261 participants in the Care Management for the Effective Use of Opioids (CAMEO) trial (NCT01236521).[18] In the CAMEO trial, patients with moderate to severe chronic low back pain were recruited from Veterans Affairs primary care clinics. The study tested the effectiveness of pharmacological (opioid management coupled with algorithm-based co-analgesic treatment) compared to behavioral approaches (education combined with pain self-management skills training) for chronic low back pain.

Sample 2 included 240 participants in the Strategies for Prescribing Analgesics Comparative Effectiveness (SPACE) trial (NCT01583985).[16] In the SPACE trial, patients with chronic back or lower extremity osteoarthritis pain with moderate-severe intensity were recruited

from primary care clinics. In this pragmatic trial, effectiveness of two flexible prescribing algorithms (opioid-avoidant versus opioid-intensive) were compared.

Sample 3 included 258 participants in the Stroke Self-Management (SSM) trial (NCT01507688). In this clinical trial, stroke survivors were recruited after being discharged from a hospital with a primary diagnosis of an acute stroke or transient ischemic attack. In this trial, the effectiveness of a stroke self-management intervention was tested against usual care. We included this stroke sample for two reasons. First, responsiveness can be context-specific;[25] therefore, evaluating a scale's responsiveness in various samples with a wide range of pain levels is essential. Second, the National Institutes of Health has called for adopting PROMIS® measures as common data elements in clinical research, patient registries, and electronic medical records to facilitate cross-study comparison and data combination.[31] As pain often accompanies other medical conditions, PROMIS-PI short forms will likely be increasingly adopted in research and clinical settings. Information about their responsiveness beyond the context of chronic musculoskeletal pain is needed.

2.2. Procedures

The Indiana University Institutional Review Board approved the study. Informed consent was obtained from all participants. Self-reported sociodemographic and clinical information was collected at baseline. At both baseline and follow-up, participants completed the four PROMIS-PI short forms, legacy measures, and a global rating of pain (described below). At follow-up, retrospective global ratings of pain change since baseline were also collected. The time frames for follow-up varied by participating studies. Specifically, follow-up measurement for the CAMEO trial was conducted at 6 months after the baseline assessment, while follow-up measurement for the SPACE and SSM trials was completed after 3 months. All study questionnaires were administered by trained research personnel.

2. 3. Measurement

2. 3.1. PROMIS-PI Short Forms.

We evaluated four fixed-length PROMIS-PI short forms: 1) the original 6-item pain interference Short Form (6b), 2) the 4-item pain interference scale (4a) included in the PROMIS-29 profile, 3) the 6 item pain interference scale (6a) included in the PROMIS-43 profile, and 4) the 8-item pain interference scale (8a) included in the PROMIS-57 profile. In 4a, 6a, and 8a short forms, the items are nested: the 6a short form was constructed by adding two items to the 4a short form and the 8a short form was constructed by adding two items to the 6a short form. The original 6-item pain interference short form (6b) shares some but not all of the items as the 4a, 6a, and 8a short forms. Among the four fixed-length PROMIS-PI short forms, there are 12 unique items. These unique items were administered at baseline and follow-up.

We selected fixed-length short forms rather than computer adaptive testing, because in many research and clinical contexts, short forms are more feasible to administer. Response format was consistent across PROMIS-PI short forms and was based on a 5-point ordinal rating scale. The response options were “Not at all,” “A little bit,” “Somewhat,” “Quite a bit,” and “Very much.” Raw score totals on each measure were converted to an item response theory-based T-score using a scoring manual (More information can be found at http://www.healthmeasures.net/images/PROMIS/manuals/PROMIS_Pain_Interference_Scoring_Manual_02232017.pdf). Following this scoring manual, we converted raw scores to T scores using the scoring tables. We calculated a scale score for an individual participant only when 50% or more of items on the short form were answered by that participant; the response scores from the answered items were summed, multiplied by the total number of items in the short form, and divided by the number of items that were answered. At the item level, the amount of missing data was minimal, ranging from 0% to 2.6% of participants across three studies. At the scale level, the amount of missing data ranged from 0% to 0.8%.

The T-score metric allows for directly comparing scores between PROMIS measures of different lengths, comparing scores between different samples, and comparing scores to the population norm.[1] The PROMIS-PI scales were calibrated in the US general population and

centered (T-score of 50 equals the mean) on a subsample that represented this reference population. This subsample was matched on important demographic variables (e.g., sex, race, and education) reported in the 2000 US General Census.[1] A higher T-score represents more of the construct being measured (i.e., more pain interference).

2.3.2. Legacy measures

2. 3.2.1. The Brief Pain Inventory Interference (BPI-I) Scale.

The BPI is among the most extensively used questionnaires in clinical pain research.[5] The 7-item BPI-I scale measures pain interference on mood, physical activity, work, social activity, relations with others, sleep, and enjoyment of life, and is conceptually comparable to the PROMIS-PI short forms. Each BPI-I item is scored 0 ("Does not interfere") to 10 ("Interferes completely"), and the BPI-I scale score is the mean score of the 7 interference items. Scores range from 0 to 10 with higher scores indicating greater pain interference. The reliability, validity, and responsiveness of the BPI are well-established.[5]

2. 3.2.2. PEG.

The PEG is a 3-item pain measure derived from the BPI.[17] The three items assess average pain intensity (P), interference with enjoyment of life (E), and interference with general activity (G). Each item is scored 0 ("no pain" or "does not interfere") to 10 ("pain as bad as you can imagine" or "interferes completely"). The PEG scale score is the mean score of the 3 items. Scores range from 0 to 10 with higher scores indicating worse pain. The PEG scale has demonstrated reliability, validity, responsiveness to change in different settings.[14; 17]

2. 3.2.3. SF Bodily Pain Scale.

The SF-36 Bodily Pain Scale was administered only in the CAMEO trial. SF-36 Bodily Pain is a 2-item subscale of the Medical Outcomes Study SF-36 questionnaire, which has been validated in different settings.[23] One item assesses pain interference on a 1 (not at all) to 5 (extremely) scale over the past four weeks. The other item assesses pain severity on a 1 (None) to 6 (very severe) scale over the past four weeks. Responses from the 2 items are summed and

then transformed to a 0–100 scale to derive a bodily pain subscale score.

The Roland-Morris Disability Questionnaire. The Roland-Morris Disability Questionnaire was administered only in the CAMEO trial. The Roland-Morris Disability Questionnaire is a checklist of 24 items assessing pain effects on function.[27] The scale score is the number of items endorsed with a possible range from 0 to 24 with higher scores indicating more pain-related disability. Substantial evidence has accumulated for the validity of the Roland-Morris Disability Questionnaire scores to discriminate levels of disability in back pain and other chronic pain conditions.[13; 23]

2.3.3. Anchor Measures

2.3.3.1. Retrospective Global Rating of Change

Our study team used the retrospective global rating of change as one anchor measure. The retrospective global rating of change assesses overall clinical response from the participant's perspective.[8] At follow-up, participants were asked to rate their pain change compared to their pain at baseline assessment. Change in pain is rated on a 7-point scale with the following response options: -3 (much worse), -2 (moderately worse), -1 (a little worse), 0 (no change), +1 (a little better), +2 (moderately better), or +3 (much better). Based on the rating, participants were further categorized into three groups, improved (+1 to +3), unchanged (0), and worsened (-1 to -3). The retrospective global rating of change has been widely used to assess responsiveness of patient-reported outcome measures.[25; 26] Literature supports its validity and clinical relevance.[8]

2.3.3.2. Prospective Global Rating of Change

Our study team used the prospective global rating of change as the second anchor measure. Specifically, at baseline and follow-up, patients were asked about their pain on average in the past 7 days (i.e., cross-sectional global rating of pain). A 5-point ordinal scale ranging from 0 = "no pain" to 4 = "very severe pain" was used.[32] To calculate the prospective global change score, we subtracted the follow-up global ratings of pain from the baseline global

ratings of pain. Change scores had a possible range of -4 to +4, where negative numbers indicated worsened pain and positive numbers improved pain. Based on the rating, participants were further categorized into three groups, improved (+1 to +4), unchanged (0), and worsened (-1 to -4).

We used this prospective anchor to overcome potential recall and reconstruction bias related to the retrospective global rating of change.[28] A few studies have suggested, compared to the retrospective global rating of change, that the prospective global rating of change may be less influenced by post-treatment status.[10; 28] Therefore, some researchers recommend it as a valid anchor for establishing true individual change.[10; 28; 32]

2.4. Data Analysis

We evaluated comparative responsiveness for all four PROMIS-PI short forms and legacy measures (i.e., BPI Pain Interference, PEG, SF-36 Bodily Pain, and Roland-Morris Disability Questionnaire). Data from each of the three trials were analyzed separately rather than pooled, because the three trials involved different clinical populations, study interventions, and follow-up timeframes.

We used both retrospective and prospective global ratings of pain change as the anchors (i.e., criteria) to identify patients who have changed since baseline. Specifically, patients were categorized into three groups based on global ratings of pain change: improved, unchanged, and worsened. Both within-group responsiveness to change and between-group responsiveness to change were evaluated as described below.

2.4.1. Within-group Responsiveness

For within-group responsiveness, we estimated the amount of change over time within each global rating of pain change group (i.e., improved, unchanged, and worsened). Standardized response means (SRM) were used as the effect size measure of within-group responsiveness to change. The SRM is the ratio of the mean change to the standardized deviation (SD) of change, and is calculated using the formula (Mean baseline – Mean follow-up)

/ SD of change score. In addition to reporting the point estimates of SRMs, we also calculated 95% confidence intervals for the SRMs with a bootstrapping procedure. An absolute SRM value of 0.2 to 0.5 is considered a small change, 0.5 to 0.8 is moderate, and ≥ 0.8 is large.[25] Some researchers suggest an absolute SRM value ≥0.3 indicates responsiveness.[2]

2.4.2. Between-Group Responsiveness

For between-group responsiveness, we compared the amount of change between global rating of change groups. First, we used omnibus ANOVA tests to compare mean change in scale scores across global rating of change groups (i.e., improved, unchanged, and worsened). For this analysis, both retrospective and prospective rating of change groups were used as anchors. We used post-hoc Tukey-Kramer tests to pair-wise compare the 1) improved and unchanged groups, and 2) worsened and unchanged groups. We controlled for family-wise Type 1 error at 0.05.

Second, we used receiver-operating characteristic curve analyses to further quantify a measure's ability to detect improvement. Area under the curve (AUC) is the probability of correctly discriminating between patients who have improved and those who have not. The AUC values range from 0.5 (the same as chance) to 1.0 (perfect discrimination). We calculated AUC for each pain measure using retrospective global rating of change as the anchor. Specifically, we evaluated each measure's ability to detect *any* improvement ("a little better", "moderately better", or "very much better") as well as *moderate* improvement ("moderately better" or "very much better"). To determine if PROMIS-PI short forms and legacy measures differ in their ability to detect improvement, we also statistically compared AUC values between PROMIS-PI and legacy measures.[6 20]

3. Results

3.1 Demographic and Clinical Characteristics

A total of 759 participants completed the baseline and follow-up assessments. Demographic characteristics of clinical trial participants have been previously published.[4] Briefly, baseline data were available on 261 participants in the CAMEO trial, 240 in the SPACE trial, and 258 in the SSM trial. CAMEO participants were 92% male, 73% white, with a mean age of 57.9 years. SPACE participants were 87% male, 86% white, with a mean age of 58.3 years. SSM participants were 81% male, 64% white, with a mean age of 61.7 years.[4]

Participants in the CAMEO and SPACE trials had greater pain interference than the US general population, reflected by their PROMIS-PI mean scores one standard deviation above the US population mean.[4] Participants in SSM trial had a pain interference level close to the US population norm.[4] The PROMIS-PI T score has a possible range of 41 to 78 .The PROMIS-PI short forms had an observed mean score of 62 (SD=7) at baseline and 60 (SD=8) at follow-up in the CAMEO trial, 62 (SD =5) at baseline and 56 (SD=7) at follow-up in the SPACE trial, and 53 (SD=10.5) at baseline and 53 (SD=10) at follow-up in the SSM trial. Ceiling effects for PROMIS-PI measures were negligible; only 0% to 7% of participants responded with the maximum possible score (the percentage varied depending on the short form, the trial, and the time point). Floor effects, which were also negligible for the pain trials (CAMEO, SPACE; 1% to 8%), were present for the stroke trial (SSM); 31% to 36% of SSM participants responded with the minimum possible score (the percentage varied depending on the short form and time point).

3.2 Within-group Responsiveness

In Figure 1, within-group effect size estimates (i.e., SRMs) were plotted for the PROMIS-PI and legacy measures across 3 trials. This figure provides an overview of comparative within-group responsiveness across the pain measures. Tables 1 and 2 complement Figure 1 by presenting the unstandardized change scores and SRM confidence intervals for the retrospective and prospective anchors, respectively.

Across the PROMIS-PI, BPI-I, and PEG measures, the SRM point estimates were

generally similar from a practical standpoint (Figures 1 and 2). In most cases, the confidence interval for one measure covered the point estimates of the other measures (Tables 1 and 2), which indirectly suggests statistically comparable within-in group responsiveness across these three measures. Minor differences in SRMs, however, were observed: For the two samples with chronic pain (i.e., CAMEO and SPACE Trials), within the pain improved group, SRMs were slightly larger for the PEG than for the PROMIS-PI and BPI-I. For the stroke sample (i.e., SSM Trial), however, SRMs for the pain worsened group were slightly larger for the PROMIS-PI scales than for the BPI-I and PEG (notably when using retrospective global change as the anchor). Based on SRMs, the SF-36-Bodily Pain scale and the Roland-Morris Disability Questionnaire appeared less responsive than PROMIS-PI, BPI-I, and PEG; however, data comparing these measures were only available for one trial.

Across the four PROMIS-PI short forms, the SRMs were statistically and practically comparable (Figure 2). The differences in SRMs between any two PROMIS-PI short forms were within 0.11. Because the SRM estimates for the four PROMIS-PI were quite similar, we reported averages of SRMs across the four PROMIS-PI short forms in Figure 1 as well as Tables 1 and 2.

For all pain measures, SRMs varied based on the samples and whether pain improved or worsened. For the two chronic pain samples with moderate to severe pain, larger effect size estimates (i.e., SRMs) were observed in the pain improved groups than in the pain worsened groups. Specifically, in the two pain samples, the effect size estimates (i.e., SRMs) for the improved group ranged from moderate to large, while SRMs for the worsened group ranged from minimal to small. For the pain worsened groups in these two pain samples, a large majority of confidence intervals for SRMs included zero. In the stroke sample (i.e., SSM Trial), by contrast, larger effect size estimates (i.e., SRM) were observed in the pain worsened groups than in the improved groups.

3.3 Between-group Responsiveness

As seen in Tables 1 and 2, all measures successfully detected differences among pain improved, unchanged, and worsened groups. Omnibus F-tests were all significant at 0.05 level, except for the Roland-Morris Disability Questionnaire when using prospective global pain of change as the anchor (p=0.06).

Figure 3 and Table 3 show the results from the AUC analysis. The PROMIS-PI short forms' probability for accurately detecting *any* improvement ranged from 0.68 to 0.72 in the two pain trials, and ranged from 0.55 to 0.59 in the stroke trial. Similarly, the accuracy of the PROMIS-PI short forms in detecting *moderate* improvement ranged from 0.70 to 0.76 for the two pain trials and at a 0.59 level for the non-pain trial.

The PROMIS-PI short forms performed similarly to legacy measures with a few exceptions: In the only pain trial that included the SF-36 Bodily Pain scale (CAMEO), the PROMIS-PI short forms had significantly higher accuracy in detecting *any* improvement and *moderate* improvement than the SF-36 Bodily Pain scale (p values ranged from 0.0003 to 0.01). In the other pain trial (SPACE), the PROMIS-PI short forms all had significantly lower accuracy in detecting *any* improvement than the BPI-I and PEG measures (p values ranged from 0.003 to 0.034). However, in detecting moderate improvement, the PROMIS PI short forms were comparable to the BPI-I and PEG.

4. Discussion

Using data from three clinical trials, we found PROMIS-PI short forms were responsive to change. Moreover, their responsiveness was largely comparable to the BPI-I and PEG measures. Consistent with Askew et al.,[2] our data indicate that the PROMIS-PI scales are responsive to global improvement. Consistent with Deyo et al.,[7] PROMIS-PI scores discriminated patients whose pain was improved, unchanged, or worsened. Previous studies have largely been observational.[2; 7; 29; 30] Our study strengthens the evidence by evaluating PROMIS-PI responsiveness in the context of three clinical trials.

This study represents a novel effort to compare responsiveness of the PROMIS-PI short forms with legacy pain measures. Kean et al. previously reported PROMIS-PI were less responsive to global change than BPI-I and PEG.[14] By contrast, we found that PROMIS-PI had generally comparable responsiveness to BPI and PEG (Figures 1 and 2). The difference between our main conclusion and Kean et al.'s may be attributable to two factors. First, Kean et al[14] used only retrospective global rating of change as the anchor. We used a prospective anchor in addition to the retrospective global rating of change anchor to overcome potential recall bias related to the retrospective measure. Second, Kean et al. included only one sample (persistent musculoskeletal pain), while in our study, responsiveness was evaluated separately in trials with three different samples.

Regarding the possibly poorer responsiveness of the SF-36 Bodily Pain and Roland-Morris Disability Questionnaire scales, it should be noted these two measures were only included in the CAMEO trial. Of note, the SF-36 Bodily Pain scale also appeared less responsive in a previous study.[15] The SF-36 Bodily Pain scale has a longer recall window (past four weeks) compared to PEG, BPI, and PROMIS-PI (past week), and this longer recall window may contribute to the lower responsiveness to change.[15] However, given the considerable use of the SF-36 Bodily Pain and Roland-Morris Disability Questionnaire scales in pain research,[11] further investigation of their comparative responsiveness is warranted.

Despite the overall comparative responsiveness across PROMIS-PI, BPI-I, and PEG, a closer inspection of our findings and Kean et al. suggest the possibility of modest performance differences in certain situations. The BPI-I and PEG may be more advantageous when patients report high pain and when detecting pain improvement is the priority (e.g., clinical trials of pain interventions/treatment). Compared to PROMIS-PI, the BPI-I and PEG had larger effect sizes for pain improvement and higher accuracy in detecting any improvement in both our SPACE trial and the trial reported by Kean et al. (both trials included patients with moderate to severe pain at baseline).[14] However, PROMIS-PI may be more advantageous when the study

population has a more heterogeneous level of pain and when both pain improvement and worsening are of interest. In our stroke trial where baseline pain was mild and no pain-specific intervention was involved, PROMIS-PI performed slightly better than BPI-I and PEG in the worsened group. The observed differences could be in part attributed to differences in the measure development processes. The BPI and PEG were developed in patients with pain using classical test theory, while PROMIS-PI were developed in both patients with pain and the general population using item response theory. As the PROMIS-PI short forms are derived from a longer item bank, it would be possible in the future to derive a short form that is better targeted and more responsive to high levels of pain. While these modest responsiveness differences across samples with higher and lower levels of pain require further investigation, our collective findings integrating both AUC and SRM analyses indicate generally comparable responsiveness of the PROMIS-PI, BPI and PEG. Importantly, even though the magnitude of SRMs differed somewhat among the scales depending upon the sample, statistical testing comparing AUCs confirmed that the scales consistently differentiated both the improved and worsened groups from the unchanged group.

This is among the few studies that used both retrospective and prospective global ratings of change as the anchors in assessing responsiveness. Retrospective global rating of change has sometimes been criticized because of its potential susceptibility to recall and reconstruction bias.[9; 28] When using retrospective global rating of change as the anchor, we observed some counter-intuitive findings. In one trial (SPACE), patients who retrospectively recalled having had worsened pain actually reported improved pain interference. Similarly, the qualitative data from the SPACE trial suggested that the participants' narrative recall of treatment effectiveness after interventions did not always match their responder status based on change in numerical scores.[22] Conversely, Prospective global rating of change which is derived by the difference in cross-sectional global ratings of pain at two time points[32] is conceptually attractive but has not been studied in depth. Thus, integrating findings from these two different

methods for assessing patient-rated global change may be preferable to relying on one method alone.

Interestingly, responsiveness varied based upon the direction of change and the particular sample. Consistent with Askew et al.[2] and Deyo et al.,[7] our data in two pain trials demonstrated a larger magnitude of change for the pain improved group than for the pain worsened groups. Specifically, for the pain improved groups, all the pain change scores had moderate-to-large effect sizes, and the magnitude of improvement in PROMIS-PI scores were above minimally important difference (2-3 points in T score[4]). For the pain worsened group, however, the PROMIS-PIs had minimal-to-small effect sizes, and the magnitude of change in PROMIS-PI scores were below minimally important difference.[4] The larger SRMs for improvement compared to worsening might be due to a couple factors. First, in these two pain trials, some patients were exposed to a pain intervention which contributed to a larger effect sizes for the improved group. Second, the two pain trials only included patients with moderate to severe pain. The restricted room for worsening may have contributed to a small magnitude of change in the pain worsening groups. Despite the ceiling effects not being substantial, participants in the two pain trials had on average only about 10 points to reach the maximum possible scores. The stroke trial participants, compared to the pain trials participants, demonstrated a larger magnitude of change in the worsened group, which is likely due to larger room for pain to worsen (about 20 points). In the stroke trial, floor effects were present with one-third participants reporting the minimum possible scores. In addition, the generally mild pain in stroke patients created a restrictive room (about 10 points on average) for their pain to improve, which may have contributed to the small magnitude of change in the improved group.

Shorter measures may be as responsive as longer measures. We found that the four PROMIS-PI versions ranging from 4 to 8 items had similar responsiveness. The PROMIS-PI short forms share some items in common, which may explain in part their comparable responsiveness. Likewise, the 3-item PEG was comparable to the 7-item BPI in our trials,

replicating findings from three previous studies.[14; 15; 20] Short measures may be more desirable for large studies with multiple outcome measures, particularly where pain may a secondary rather than the primary outcome, or in busy clinical practice settings.

This study has several strengths. First, we compared the responsiveness of PROMIS-PI short forms with widely used legacy pain measures. Second, we used both retrospective and prospective global rating of change anchors. Third, we tested responsiveness in the context of clinical trials. Fourth, we evaluated responsiveness to change using three clinical samples, each of which was large enough for psychometric evaluation. Lastly, we evaluated responsiveness in both directions of change (improvement and worsening).

Our study has several limitations. First, two legacy measures (SF-36 Bodily Pain and Roland-Morris Disability Questionnaire) were only used in one trial (CAMEO). We found PROMIS appeared superior to these two measures in responsiveness to change; however, generalizability of this finding may be limited. Second, our two pain samples had at least moderate pain intensity at baseline, leaving a limited room for pain to worsen as discussed earlier. Third, we made multiple statistical comparisons between multiple pain measures. Readers need to interpret the differences between measures with caution. Fourth, our retrospective anchor may reflect a change in pain intensity and/or interference. We selected this anchor based on the existing literature; [2; 14; 25; 32] however, it is unknown how participants interpreted this anchor item. If participants focused on perceived change on pain intensity rather than pain interference, the conceptual difference between pain intensity and interference may partially explain some counter-intuitive findings. Nonetheless, pain intensity and interference tend to respond in parallel when used as outcome measures in clinical trials.[15; 19]

Despite these limitations, this study has several implications for research and clinical practice. First, either the PROMIS-PI or legacy measures like BPI or PEG are reasonable choices based on responsiveness to change. As the use of the BPI requires a permission from the developers and in some circumstances a fee, the freely-available PROMIS-PIs and PEG

can be acceptable alternatives. Second, future research is needed to evaluate the comparative responsiveness of the PROMIS-PI scales in additional samples. Although this study provided support for PROMIS-PI responsiveness, generalizability to other pain populations (e.g., acute, recurrent, visceral, neuropathic pain) should be studied. Third, researchers need to further evaluate appropriateness of various anchors in assessing responsiveness. The counter-intuitive finding regarding retrospective global rating of change we found in one trial was also reported in another study.[2] Thus, including both prospective and retrospective global rating of change anchors may allow researchers to check the robustness of findings.

Acknowledgements

This work was supported by a National Institute of Arthritis and Musculoskeletal Disorders R01 award to Dr. Monahan (R01 AR064081) and Department of Veterans Affairs Health Services Research and Development Merit Review awards to Drs. Bair (IIR 10-128), Krebs (IIR 11-125), and Damush (VA HSRD QUERI Service Directed Project SDP- 10-379). Dr. Chen was supported by the National Institute of Nursing Research under award number 5T32 NR007066, the Indiana University–Purdue University Indianapolis Enhanced Mentoring Program with Opportunities for Ways to Excel in Research (EMPOWER) Grant, and Grants Numbers KL2TR001106 and UL1TR001108 (Shekhar, PI) funded by the National Institutes of Health, National Center for Advancing Translational Sciences Clinical and Translational Sciences Award. Dr. Kean was supported by the Department of Veterans Affairs Rehabilitation Research and Development Career Development Award (IK2RX000879). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the Department of Veteran Affairs. The authors thank Dr. Janet Carpenter for helpful comments on earlier versions of this paper and Ms. Lindsay Rosa for the editorial support.

Conflict of Interest

The authors declare no conflict of interest.

References

References

[1] Amtmann D, Cook KF, Jensen MP, Chen WH, Choi S, Revicki D, Cella D, Rothrock N, Keefe F, Callahan L, Lai JS. Development of a PROMIS item bank to measure pain interference. Pain 150:173-182, 2010.

[2] Askew RL, Cook KF, Revicki DA, Cella D, Amtmann D. Evidence from diverse clinical populations supported clinical validity of PROMIS pain interference and pain behavior. J Clin Epidemiol 73:103-111, 2016.

[3] Cella D, Yount S, Rothrock N, Gershon R, Cook K, Reeve B, Ader D, Fries JF, Bruce B, Rose M. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. Med Care 45:S3-S11, 2007.

[4] Chen CX, Kroenke K, Stump TE, Kean J, Carpenter JS, Krebs EE, Bair MJ, Damush TM, Monahan PO. Estimating minimally important differences for the PROMIS pain interference scales: results from 3 randomized clinical trials. Pain 2018;159:775-782.

[5] Cleeland C, Ryan K. Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singapore 23:129-138, 1994.

[6] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837-845, 1988.

[7] Deyo RA, Ramsey K, Buckley DI, Michaels L, Kobus A, Eckstrom E, Forro V, Morris C. Performance of a Patient Reported Outcomes Measurement Information System (PROMIS) Short Form in Older Adults with Chronic Musculoskeletal Pain. Pain Med 17:314-324, 2015.

[8] Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, Stucki G, Allen RR, Bellamy N, Carr DB, Chandler J, Cowan P, Dionne R, Galer BS, Hertz S, Jadad AR, Kramer LD, Manning DC, Martin S, McCormick CG, McDermott MP, McGrath P, Quessy S, Rappaport BA, Robbins W, Robinson JP, Rothman M, Royal MA, Simon L, Stauffer JW, Stein W, Tollett J, Wernicke J, Witter J. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain 113:9-19, 2005.

[9] Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT, Haythornthwaite JA, Jensen MP, Kerns RD, Ader DN, Brandenburg N, Burke LB, Cella D, Chandler J, Cowan P, Dimitrova R, Dionne R, Hertz S, Jadad AR, Katz NP, Kehlet H, Kramer LD, Manning DC, McCormick C, McDermott MP, McQuay HJ, Patel S, Porter L, Quessy S, Rappaport BA, Rauschkolb C, Revickl DA, Rothman M, Schmader KE, Stacey BR, Stauffer JW, Von Stein T, White RE, Witter J, Zavislc S. Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 9:105-121, 2008.

[10] Fletcher KE, French CT, Irwin RS, Corapi KM, Norman GR. A prospective global measure, the Punum Ladder, provides more valid assessments of quality of life than a retrospective transition measure. J Clin Epidemiol 63:1123-1131, 2010.

[11] Goldsmith ES TB, Greer N, Murdoch M, MacDonald R, McKenzie L, Rosebush CE, Wilt TJ. Focused evidence review: psychometric properties of patient-reported outcome measures for chronic musculoskeletal pain. J Gen Intern Med 33:61-70, 2018.

[12] Hinchcliff ME, Beaumont JL, Carns MA, Podlusky S, Thavarajah K, Varga J, Cella D, Chang RW. Longitudinal evaluation of PROMIS-29 and FACIT-dyspnea short forms in systemic sclerosis. J Rheumatol 42:64-72. 2015.

[13] Jensen MP, Strom SE, Turner JA, Romano JM. Validity of the Sickness Impact Profile Roland scale as a measure of dysfunction in chronic pain patients. Pain 50:157-162, 1992.

[14] Kean J, Monahan PO, Kroenke K, Wu J, Yu Z, Stump TE, Krebs EE. Comparative Responsiveness of the PROMIS Pain Interference Short Forms, Brief Pain Inventory, PEG, and SF-36 Bodily Pain Subscale. Med Care 54:414-421, 2016.

[15] Krebs EE, Bair MJ, Damush TM, Tu W, Wu J, Kroenke K. Comparative responsiveness of pain outcome measures among primary care patients with musculoskeletal pain. Med Care 48:1007-1014, 2010.

[16] Krebs EE, Jensen AC, Nugent S, DeRonne B, Rutks I, Leverty D, Gravely A, Noorbaloochi S, Bair MJ, Kroenke K. Design, recruitment outcomes, and sample characteristics of the Strategies for Prescribing Analgesics Comparative Effectiveness (SPACE) trial. Contemp Clin Trials 62:130-139, 2017.

[17] Krebs EE, Lorenz KA, Bair MJ, Damush TM, Wu J, Sutherland JM, Asch SM, Kroenke K. Development and Initial Validation of the PEG, a Three-item Scale Assessing Pain Intensity and Interference. J Gen Intern Med 24:733-738, 2009.

[18] Kroenke K, Evans E, Weitlauf S, McCalley S, Porter B, Williams T, Baye F, Lourens SG, Matthias MS, Bair MJ. Comprehensive vs. Assisted Management of Mood and Pain Symptoms (CAMMPS) trial: Study design and sample characteristics. Contemp Clin Trials 64:179-187, 2018.

[19] Kroenke K, Krebs EE, Wu J, Yu Z, Chumbler NR, Bair MJ. Telecare collaborative management of chronic pain in primary care: a randomized clinical trial. JAMA 312:240-248, 2014.

[20] Kroenke K, Theobald D, Wu J, Tu W, Krebs EE. Comparative responsiveness of pain measures in cancer patients. J Pain 13:764-772, 2012.

[21] Lee AC, Driban JB, Price LL, Harvey WF, Rodday AM, Wang C. Responsiveness and Minimally Important Differences for 4 Patient-Reported Outcomes Measurement Information System Short Forms: Physical Function, Pain Interference, Depression, and Anxiety in Knee Osteoarthritis. J Pain 18:1096-1110, 2017.

[22] Matthias MS, Donaldson MT, Jensen AC, Krebs EE. "I Was a Little Surprised": Qualitative Insights From Patients Enrolled in a 12-Month Trial Comparing Opioids With Nonopioid Medications for Chronic Musculoskeletal Pain. J Pain 19: 1082-1091, 2018.

[23] McHorney CA, Ware JE, Jr., Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 31:247-263, 1993.

[24] Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, Bouter LM, de Vet HC. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol 63:737-745, 2010.

[25] Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol 61:102-109, 2008.

[26] Roland M, Fairbank J. The Roland-Morris Disability Questionnaire and the Oswestry Disability Questionnaire. Spine (Phila Pa 1976) 25:3115-3124, 2000.

[27] Roland M, Morris R. A study of the natural history of back pain. Part I: development of a reliable and sensitive measure of disability in low-back pain. Spine (Phila Pa 1976) 8:141-144, 1983.

[28] Schmitt J, Di Fabio RP. The validity of prospective and retrospective global change criterion measures. Arch Phys Med Rehabil 86:2270-2276, 2005.

[29] Shahgholi L, Yost KJ, Carter RE, Geske JR, Hagen CE, Amrami KK, Diehn FE, Kaufmann TJ, Morris JM, Murthy NS, Wald JT, Thielen KR, Kallmes DF, Maus TP. Correlation of the Patient Reported Outcomes Measurement Information System with legacy outcomes measures in assessment of response to lumbar transforaminal epidural steroid injections. AJNR Am J Neuroradiol 36:594-599, 2015.

[30] Shahgholi L, Yost KJ, Kallmes DF. Correlation of the National Institutes of Health Patient Reported Outcomes Measurement Information System Scales and Standard Pain and Functional Outcomes in Spine Augmentation. AJNR Am J Neuroradiol 33:2186-2190, 2012.

[31] Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, Lang L, Moser RP, Odenkirchen J, Reeves D, Rubinstein Y, Werner E, Huerta M. Improving the value of clinical research through the use of Common Data Elements. Clinical Trials 13:671-676, 2016.

[32] Yost KJ, Eton DT, Garcia SF, Cella D. Minimally important differences were estimated for six Patient-Reported Outcomes Measurement Information System-Cancer scales in advanced-stage cancer patients. J Clin Epidemiol 64:507-516, 2011.

Figure 1. Comparative Standardized Response Means (SRMs) Between Measures across Trials

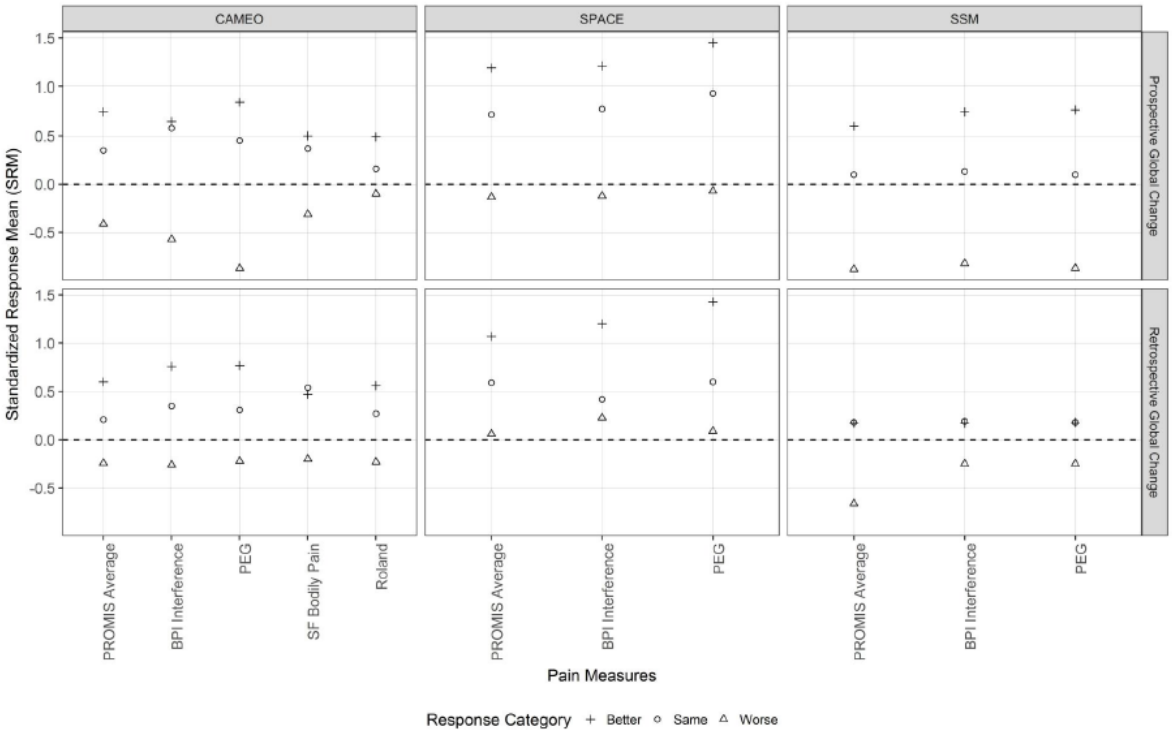

Figure 2. Comparative Standardized Response Means (SRMs) Between PROMIS Pain Interference Short Forms of Varying Lengths

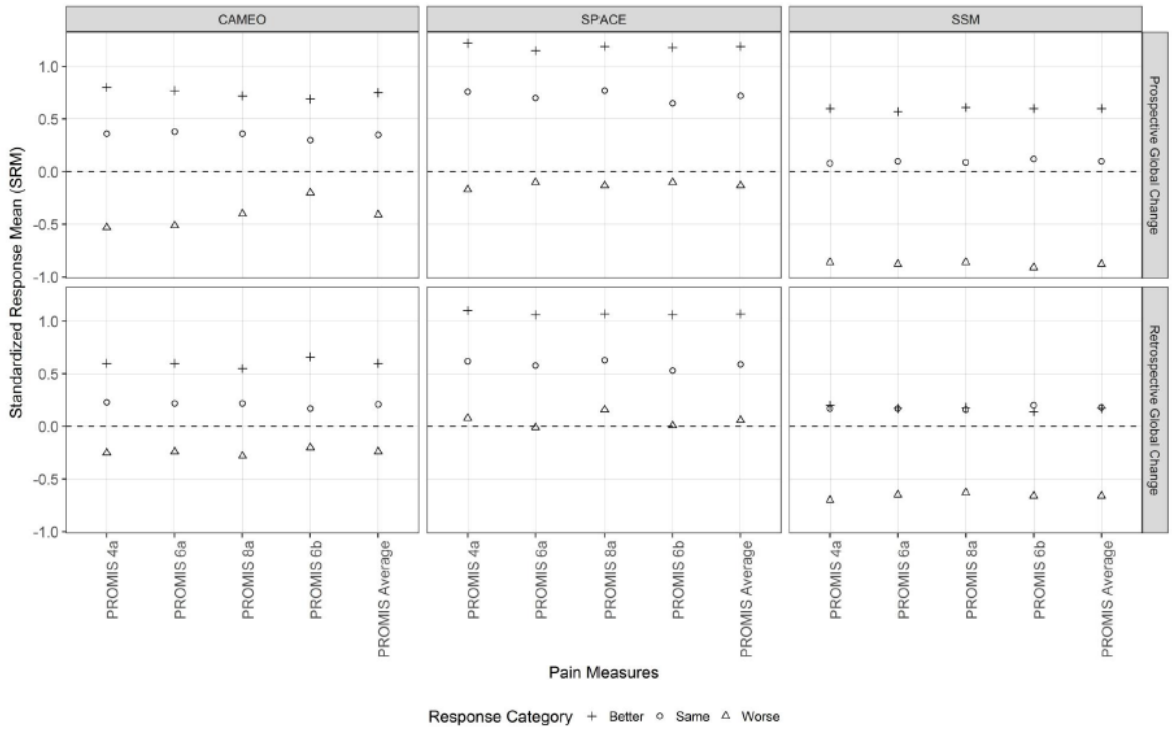

Figure 3. Comparative Area under the Curve (i.e., Accuracy) for Detecting Pain Improvement

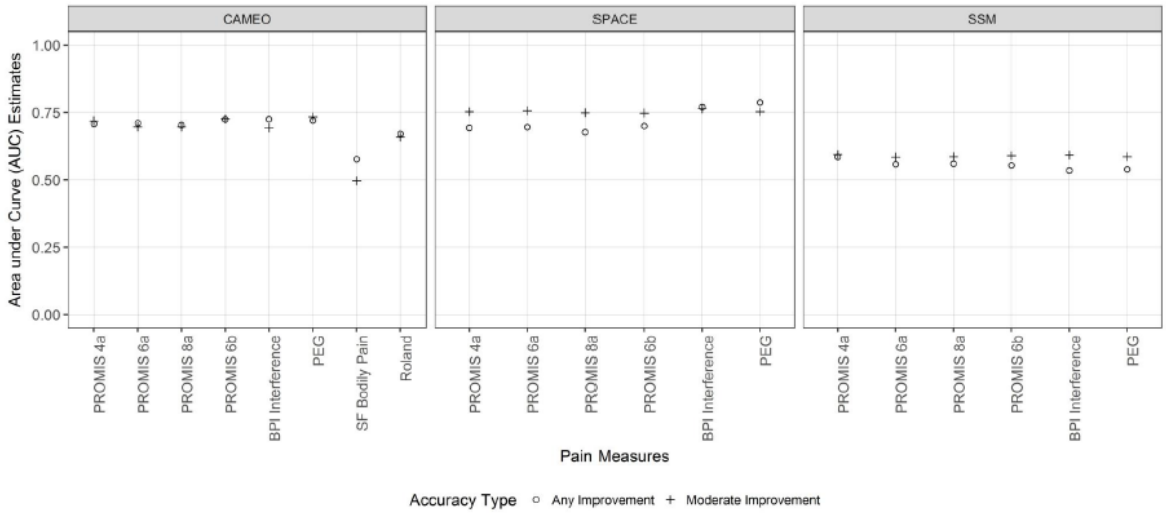

Table 1. Responsiveness by Retrospective Global Pain of Change[a]

		CAMEO Trial			SPACE Trial			SSM Trial			
Pain change	**MeanSRM**	**Score[b] change**	**SRM[c]**	**(95% CI)**	**P-value[d]**	**Score[b] change**	**SRM[c]**	**(95% CI)**	**P-value[d]**	**Score[b] change**	**SRM[c]**	**(95% CI)**	**P-value[d]**
PROMIS 4a					**<.0001**				**<.0001**				**<.0001**
• Better	.63	4.52	.60	(.40, .81)	.003	6.84	1.10	(.93, 1.13)	.0002	1.92	.20	(.01, .40)	.94
• Same	.34	1.10	.23	(-.01, .46)	--	3.27	.62	(.37, .90)	--	1.47	.17	(-.03, .36)	--
• Worse	-.30	-1.52	−.25	(-.53, .01)	.066	0.21	.08	(-.35, .59)	.089	-6.11	−.70	(-.99, -.43)	<.0001
PROMIS 6a					**<.0001**				**<.0001**				**<.0001**
• Better	.61	4.66	.60	(.39, .83)	.003	6.92	1.06	(.92, 1.2)	<.0001	1.68	.17	(-.02, .36)	.99
• Same	.32	1.13	.22	(−.01, .44)	--	3.05	.58	(.31,.88)	--	1.50	.17	(-.03, .36)	--
• Worse	-.30	-1.45	−.24	(-.53, .02)	.082	-0.07	−.01	(-.46, .43)	.094	-5.81	−.65	(-.92, -.39)	.0001
PROMIS 8a					**<.0001**				**<.0001**				**<.0001**
• Better	.60	4.17	.55	(.35, .79)	.012	6.69	1.07	(.92, 1.25)	.0002	1.72	.18	(-.01, .37)	.98
• Same	.34	1.15	.22	(-.01, .44)	--	3.17	.63	(.36, .95)		1.44	.16	(-.03, .36)	--
• Worse	-.25	-1.71	−.28	(-.57, -.01)	.042	0.63	.16	(-.27, .65)	.182	-5.54	−.63	(-.87, -.39)	.0002
PROMIS 6b					**<.0001**				**<.0001**				**<.0001**
• Better	.62	4.45	.66	(.46,	.0006	6.02	1.06	(.92, 1.22)	.0002	1.36	.14	(-.06,	.96

			.87)							.33)	
• Same	.30	0.88	.17 (-.06, .39)	--	2.67	.53 (.30, .77)	--	1.71	.20 (-.01, .39)	--
• Worse	-.28	-1.10	−.20 (-.46, .06)	.159	-0.06	.01 (-.40, .53)	.103	-5.93	−.66 (-.91, -.42)	<.0001
PROMIS average										
• Better	.62		.60			1.07			.17	
• Same	.33		.21			.59			.18	
• Worse	-.28		−.24			.06			−.66	
BPI Interference				**<.0001**			**<.0001**			**.024**
• Better	.71	1.77	.76 (.56, .97)	.0003	2.51	1.20 (1.02, 1.4)	<.0001	0.43	.17 (-.02, .36)	.999
• Same	.32	0.51	.35 (.09, .61)	--	0.64	.42 (.18, .68)	--	0.41	.19 (-.01, .38)	--
• Worse	-.18	-0.45	−.26 (-.54, .00)	.019	0.32	.23 (-.21, .69)	.78	-0.80	−.25 (-.56, .06)	.033
PEG				**<.0001**			**<.0001**			**.034**
• Better	.79	1.67	.77 (.57, .99)	.0005	2.61	1.43 (1.25, 1.6)	<.0001	0.46	.18 (-.01, .37)	.994
• Same	.36	0.53	.31 (.08, .54)	--	0.86	.60 (.35, .87)	--	0.42	.18 (-.02, .39)	--
• Worse	-.13	-0.34	−.22 (-.49, .04)	.028	0.14	.09 (-.34, .57)	.203	-0.73	−.25 (-.57, .06)	.048
SF Bodily Pain[d]				**<.0001**						

• Better		8.13	.47	(.29, .66)	.838						
• Same		6.71	.54	(..29, .79)	--						
• Worse		-3.50	−.20	(-.45, .04)	.001						
Roland-Morris[d]					**<.0001**						
• Better		2.90	.56	(.40, .73)	.019						
• Same		1.01	.27	(.05, .49)	--						
• Worse		-0.84	−.23	(-.49, .02)	.048						

[a]Total N (better, same, worse) with baseline and follow-up data in CAMEO = 234 (104, 68, 62); in SPACE = 222 (134, 66, 22); and in SSM = 238 (102, 96, 40)
[b]Score change = baseline - follow-up (Positive score: improvement, negative score: worsening)
[c] SRM = baseline – follow-up / SD change score;

[d] Bolded p-values are from omnibus ANOVA tests comparing changes scores among three change groups. Other p values were derived from t-test comparing change scores between reference (i.e. “same”) and changed (“better” or “worse”) groups and were adjusted for multiple comparison

Table 2. Responsiveness by Prospective Global Pain of Change[a]

		CAMEO Trial				SPACE Trial				SSSM Trial			
Pain change	**MeanSRM**	**Score[b] change**	**SRM[c]**	**(95% CI)**	**P-value[d]**	**Score[b] change**	**SRM[c]**	**(95% CI)**	**P-value[d]**	**Score[b]change**	**SRM[c]**	**(95% CI)**	**P-value[d]**
PROMIS-PI 4a					**.0007**				**<.0001**				**<.0001**
• Better	.87	5.31	.80	(.44, 1.2)	.111	8.11	1.22	(0.98, 1.5)	<.0001	5.64	.60	(.40, .80)	<.0001
• Same	.40	2.57	.36	(.15, .57)	--	3.45	.76	(.59, .95)	--	0.47	.08	(-.12, .28)	--
• Worse	-.52	-2.39	-.53	(-.98, -.12)	.018	-0.82	-.17	(-.73, .35)	.014	-7.68	-.86	(-1.2, -.59)	<.0001
PROMIS-PI 6a					**.004**				**<.0001**				**<.0001**
• Better	.83	4.97	.77	(.35, 1.3)	.277	8.06	1.15	(.93, 1.4)	<.0001	5.47	.57	(.38, .76)	.0003
• Same	.39	2.83	.38	(.16, .60)	--	3.33	.70	(.52, .89)	--	0.61	.10	(-.10, .29)	--
• Worse	-.50	-1.95	-.51	(-.11, -.04)	.028	-0.41	-.10	(-.74, .40)	.052	−7.88	-.88	(-1.2, -.62)	<.0001
PROMIS-PI 8a					**.001**				**<.0001**				**<.0001**
• Better	.87	4.58	.72	(.34, 1.1)	.287	7.81	1.19	(.95, 1.5)	<.0001	5.68	.61	(.43, .80)	<.0001
• Same	.41	2.54	.36	(.14, .59)	--	3.49	.77	(.58, .97)	--	0.51	.09	(-.11, .28)	--
• Worse	-.46	-1.46	-.40	(-.92, .08)	.066	-0.51	-.13	(-.76, .37)	.024	−7.84	-.86	(-1.2, -.59)	<.0001

PROMIS 6b					**.006**				**<.0001**			**<.0001**
• Better	.82	4.75	.69	(.35, 1.1)	.053	7.04	1.18	(1.00, 1.4)	<.0001	5.57	.60	(.42, .78)	.0003
• Same	.36	1.84	.30	(.08, .52)	--	2.86	.65	(.47, .85)	--	0.71	.12	(-.08, .31)	--
• Worse	-.40	-0.74	-.20	(-.76, .29)	.26	-0.41	-.10	(-.74, .39)	.056	−8.51	-.91	(-1.2, -.66)	<.0001
PROMIS average													
• Better	.85		.75				1.19				.60		
• Same	.39		.35				.72				.10		
• Worse	-.47		-.41				-.13				-.88		
BPI Interference					**<.0001**				**<.0001**			**<.0001**
• Better	.87	1.66	.65	(.35, .97)	.119	2.70	1.21	(.99, 1.4)	<.0001	1.81	.75	(.58, .93)	<.0001
• Same	.50	0.89	.58	(.35, .83)	--	1.20	.78	(.58, .99)	--	0.19	.13	(-.07, .33)	--
• Worse	-.50	-1.06	-.57	(-1.2, -.09)	.0007	-0.13	-.12	(-.83, .38)	.027	−2.13	-.82	(-1.1, -.57)	<.0001
PEG					**<.0001**				**<.0001**				**<.0001**
• Better	1.02	2.02	.85	(.60, 1.1)	.0007	2.85	1.45	(1.23, 1.7)	<.0001	1.93	.77	(.59, .97)	<.0001
• Same	.50	0.67	.45	(.24, .67)	--	1.30	.94	(.73, 1.2)	--	0.15	.10	(-.09, .30)	--
• Worse	-.60	-1.22	-.87	(-1.6	.0003	-0.13	-.07	(-.62,	.007	-2.11	-.87	(-1.1	<.0001

				, -.34)			.45)			, -.64)	
SF Bodily Pain[ll]					**.013**						
• Better		10.18	.50	(.20, .80)	.296						
• Same		5.14	.37	(.12, .63)	--						
• Worse		-4.44	-.31	(-1.1, .20)	.084						
Roland-Morris[ll]					**.055**						
• Better		2.18	.49	(.18, .83)	.138						
• Same		0.62	.16	(-.05, .38)	--						
• Worse		-0.44	-.10	(-.59, .40)	.59						

[a] Total N (better, same, worse) with baseline and follow-up data in CAMEO = 135 (38, 79, 18); in SPACE = 222 (94, 112, 16); and in SSM = 238 (83, 100, 55);
[b]Score change = baseline - follow-up (positive score: improvement, negative score: worsening);
[c] SRM = baseline – follow-up / SD change score;
[d]bolded P-values are from omnibus ANOVA tests comparing changes scores among 3global ratings of change groups. Other P-values are derived from t-test comparing change scores between reference (i.e. “same”) and changed (“better” or “worse”) groups and were adjusted for multiple comparison

Table 3. Area under the Curve (AUC) for Pain Measures in 3 Trials[a]

Pain Scale	Average accuracy across trials		Accuracy for detecting any improvement[b]						Accuracy for detecting moderate improvement[b]					
	Any improvement	Moder-ate improvement	CAMEO		SPACE		SSM		CAMEO		SPACE		SSM	
			AUC	(95% CI)	AUC	(95% CI)	AUC	(95% CI)	AUC	(95% CI)	AUC	(95% CI)	AUC	(95% CI)
PROMIS 4a	.662	.689	.708	(.640-.775)	.693	(.624-.763)	.585	(.513-.657)	.719	(.638-.799)	.753	(.686-.819)	.594	(.518-.670)
PROMIS 6a	.655	.679	.711	(.643-.779)	.696	(.627-.766)	.558	(.485-.631)	.697	(.609-.784)	.756	(.690-.821)	.585	(.508-.662)
PROMIS 8a	.647	.677	.704	(.635-.773)	.677	(.607-.748)	.560	(.487-.633)	.697	(.607-.786)	.749	(.682-.817)	.586	(.509-..663)
PROMIS 6b	.659	.688	.724	(.657-.791)	.700	(.630-.769)	.554	(.480-.627)	.726	(.641-.811)	.748	(.680-.815)	.590	(.513-.668)
PROMIS average	.656	.683	.712		.692		.564		.710		.752		.589	
BPI Interference	.677	.683	.725	(.660-.790)	.770[d]	(.710-.830)	.535	(.464-.606)	.694	(.604-.784)	.764	(.702-.826)	.592	(.519-.665)
PEG	.682	.691	.720	(.654-.785)	.787[d]	(.728-.845)	.539	(.467-.611)	.733	(.647-.819)	.753	(.691-.816)	.586	(.511-.661)
SF Bodily Pain			.577[c]	(.503-.650)	--	--	--	--	.497[c]	(.393-.601)	--	--	--	--
Roland-			.671	(.602-.740)	--	--	--	--	.659	(.563-	--	--	--	--

Morris						.755)		

[a]Total N (better, same, worse) with baseline and follow-up data in CAMEO = 234 (104, 68, 62); in SPACE = 222 (134, 66, 22); and in SSM = 238 (102, 96, 40)

[b]Any improvement ≥ “a little better”; moderate improvement ≥ “moderately better”; The AUC’s of the PROMIS Pain Interference short forms were mostly comparable to legacy pain measures with a few exceptions (below)

[c] In CAMEO trial, the SF Bodily Pain had significantly lower accuracy in detecting any improvement and moderate improvement than the other scales (p values ranged from 0.0003 to 0.01).

[d] In SPACE trial, the BPI Interference and PEG had significantly higher accuracy in detecting any improvement than the PROMIS-PI short forms.